Distributed SPARQL Query Processing enabling Virtual Data Integration for E-Science Grids

Andreas Langegger

Institute of Applied Knowledge Processing
Johannes Kepler University Linz, Austria
http://www.faw.at
al@jku.at

1 Introduction

For scientific collaboration, sharing data between different parties is fundamental. Grids, originally developed for high-performance and parallel computing, enable the sharing of distributed resources across institutional boundaries by providing a security infrastructure and standardized Grid-services. Because data is usually stored in different information systems and schemes, at the moment they have to be prepared and manually aligned to a common schema. Knowledge about data structures and semantics is a precondition to be able to integrate data sources. To enable virtual integration, several concepts have been proposed in the field of distributed and federated database systems. For the integration of heterogeneous information systems, the mediator-wrapper architecture can be used. In order to fulfill the requirements of a Grid-based data integration middleware for distributed, heterogeneous data sources, several concepts introduced in the Semantic Web community have been considered. The Resource Description Framework (RDF) is well suited for global schema management. It is simple, supports modularization of commonly used semantics by the ontology layer, and allows for reasoning. A standardized query language (SPARQL) is currently being developed.

2 Related Work and Proposed Approach

The use of ontologies for data integration is not new [1]. However, there is currently no approach which enables the integration of distributed, heterogeneous data sources by a SPARQL query processor. Related work can be divided into several domains: RDF triple stores [2, 7], RDF-based query algebra and processing [4], schema mapping, data integration [6], as well as distributed query processing [5, 8]. For some specific wrappers existing mapping frameworks can be applied, like for example D2R-Map [3] for relational database systems.

At the moment RDF and the other Semantic Web layers are not well suited for virtual data access. Although SPARQL provides a communication protocol, queries are executed on local sites only. In this PhD thesis, a new approach will be proposed to support query planning and execution in a distributed environment.

Global queries are processed by a mediator, which computes the optimal query plan by iterative dynamic programming. Wrappers for different information systems provide specific access and data manipulation functions. Depending on wrapper capabilities, multiple-set operations (e.g. join) can be executed locally or at the mediator. An iterator-based approach and concepts like row blocking, semi-joins, etc. are desirable to improve query processing performance.

The middleware is being developed within the Austrian Grid Project. There is also tight cooperation with several application workpackages. One of the prototype applications will be a *Virtual Observatory* for solar phenomenons developed together with the Kanzelhöhe Solar Observatory.

3 Outlook

Currently, there is no approach for virtual data integration based on systematic SPARQL query processing. Queries are either executed locally or targeted against single sites. Within this PhD thesis, a query processor will be developed based on the mediator-wrapper architecture, enabling virtual integration of heterogeneous, distributed data sources. The impact and sustainability is expected to be high in future.

Acknowledgements This work is supported by the *Austrian Grid Project*, funded by the Austrian *Federal Ministry for Education, Science and Culture* under contract GZ 4003/2-VI/4c/2004.

References

1. Vladimir Alexiev, Michael Breu, Jos de Bruijn, Dieter Fensel, Ruben Lara, and Holger Lausen, editors. *Information Integration with Ontologies: Ontology Based Information Integration in an Industrial Setting.* Wiley & Sons, 2005.
2. J. Carroll, I. Dickinson, C. Dollin, D. Reynolds, A. Seaborne, and K. Wilkinson. Jena: Implementing the Semantic Web Recommendations. In *Proceedings of the International World Wide Web Conference*, page 74. Hewlett Packard Labs, 2004.
3. Chris Bizer and Richard Cyganiak. D2R Server – Publishing Relational Databases on the Semantic Web. In *5th International Semantic Web Conference*, 2006.
4. Richard Cyganiak. A relational algebra for SPARQL. Technical Report HPL-2005-170, HP Labs, Bristol, UK, 2005.
5. Donald Kossmann. The state of the art in distributed query processing. *ACM Comput. Surv.*, 32(4):422–469, 2000.
6. Gio Wiederhold. Mediators in the Architecture of Future Information Systems. In A.R.Hurson, M.W.Bright, and S.H.Pakzad, editors, *Multi-database Systems: An Advanced Solution for Global Information Sharing.* IEEE Press, 1993.
7. S. Harris and N. Gibbins. 3store: Efficient Bulk RDF Storage. In *Proceedings of the First International Workshop on Practical and Scalable Semantic Systems*, Oct 2003.
8. M. Nedim Alpdemir, Arijit Mukherjee, Anastasios Gounaris, Norman W. Paton, Paul Watson, Alvaro Fernandes, and Jim Smith. OGSA-DQP: A Service-Based Distributed Query Processor For The Grid. In *Proceedings of the Second e-Science All Hands Meeting*, 2003.